towards an algorithm for

extracting exciting enzymes from metagenomic data

Samuel M. Nicholls[1,2] Wayne Aubrey[1] Kurt de Grave[2]
Leander Schietgat[2] Christopher J. Creevey[1] Amanda Clare[1]

[1]Aberystwyth University, [2]Katholieke Universiteit Leuven

Abstract

High-throughput sequencing has allowed us to look beyond consensus sequences to the variation observed within organisms; their ***haplotypes***.

However, existing approaches for recovery of haplotypes make assumptions that are violated when investigating sequences that originate from communities of microbes: ***metagenomes***.

We present ***Hansel*** and ***Gretel***: a data structure and algorithm that form a framework for the recovery of haplotypes from metagenomes.

Our approach does not require parameters or *a priori* knowledge, makes no assumptions of allelic distribution, does not need to distinguish error from variation and uses all evidence.

Exciting Exploitable Enzymes

- Members of microbial communities have adapted to produce enzymes to fulfil a niche in their environment.
- If isolated, these enzymes could be exploited in a wealth of scenarios including the refinement of biofuels, production of plastics, decontamination of polluted air and water or even the creation of new antibiotics.
- Yet it is currently difficult to isolate the producers of these enzymes from their environment by culture. We instead turn to mass environmental sequencing: ***metagenomics***.

Metagenomes and Pseudo-References

- When assembling reads that originate from more than one species, the resulting assembly will be chimeric.
- A metagenomic assembly can be a ***pseudo-reference*** to which we align both raw reads and target genes.
- This effectively screens raw reads against sequences of interest (such as enzymes) in the metagenome.

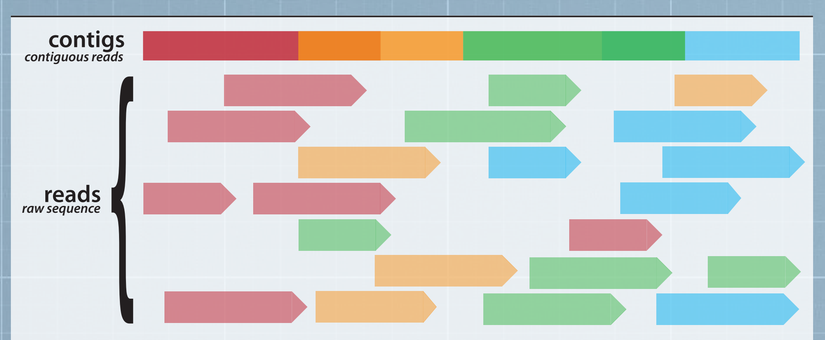

Fig 1 Overlapping reads from multiple species (coloured) can and will overlap with each-other causing constructed contigs to be chimeric.

The Problem: the Metahaplome

- Consensus sequences pose a problem for gene recovery, sequences derived from metagenomic assemblies are unlikely to actually exist in nature.
- We can't arbitrarily synthesize DNA sequences from a metagenomic assembly and expect an enzyme to work.
- **We must recover the actual haplotypes for the gene.**

We define the ***metahaplome*** as the set of haplotypes that exist for any particular genomic region of interest within a metagenomic data set and introduce a framework to recover haplotypes from such metahaplomes.

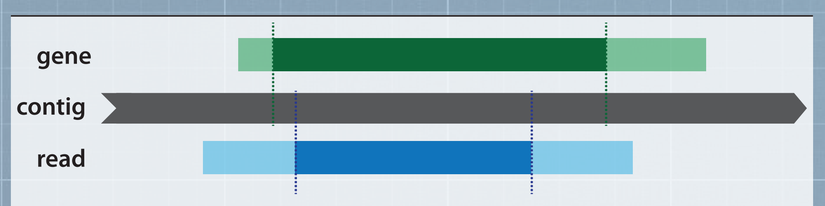

Fig 2 A gene of interest may be aligned to an assembled contig. Individual reads that then "map back" to the same region are a proxy hit on that gene. We can now focus investigation on this specific region.

Introducing the Metahaplome

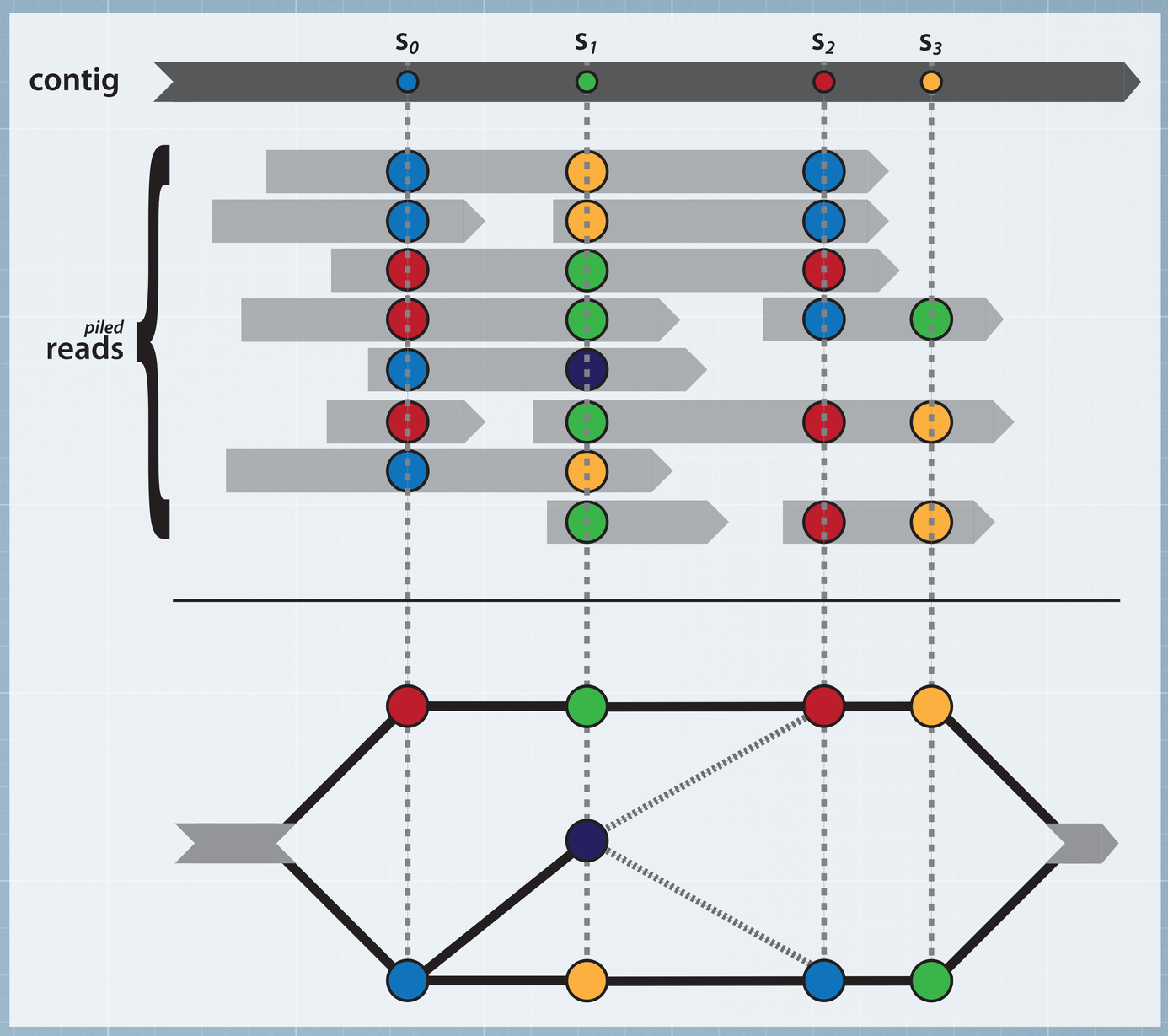

Fig 3 Above, a small set of reads from a metagenomic sample are stacked according to their alignment to an assembled contig. Sites s_0 to s_3 are SNPs that have been called by a pipeline. Bases (or "symbols") are represented as arbitrary colours. Beneath, variation in the reads is modelled by a graph. Nodes are variants at corresponding sites, and an edge between nodes exists if at least one read in the dataset demonstrates that pair of variants. The graph encodes adjacent pairwise variation observed across the reads and a path through the graph represents a haplotype in the metahaplome.

A First Model

Variants on aligned reads are parsed into a structure that stores the evidence for one variant co-occurring with another on the same read, and metadata such as quality. This evidence is exploited to build a graph, a path through which represents a single underlying haplotype.

To find likely haplotypes, we weight graph edges probabilistically, but calculating conditionals is expensive and impractical even for reasonably sized graphs. Naive Bayes offers a simple method for estimating conditional probabilities at the cost of its naive assumptions (that often prove robust in practice).

$$\mathbb{P}(v_{i+1} \mid v_{i-L}, \ldots, v_{i-2}, v_{i-1}, v_i) = \mathbb{P}(v_{i+1}) \prod_{l=0}^{L} \mathbb{P}(v_{i-l} \mid v_{i+1})$$

Fig 4 ▲ Predicting the next variant v_{i+1} given the previous L seen variants.
▼ Application of the naive assumption to predict ● following ●●.

$$\mathbb{P}(\bullet \mid \bullet\bullet) = \mathbb{P}(\bullet_{i+1})\,\mathbb{P}(\bullet_{i-1} \mid \bullet_{i+1})\,\mathbb{P}(\bullet_{i} \mid \bullet_{i+1}) = \mathbb{P}(\bullet_{3})\;\mathbb{P}(\bullet_{1} \mid \bullet_{3})\;\mathbb{P}(\bullet_{2} \mid \bullet_{3})$$

We are able to quickly and cheaply estimate the probability of a particular variant occurring given those observed on the path (sequence of SNPs) so far.

$$\mathbb{P}(\bullet_2 \mid \bullet_3) = \frac{1 + (\#\text{Reads SNP[2]}=\bullet \;\&\; \text{SNP[3]}=\bullet)}{\#\text{Variants SNP[2]} + (\#\text{Reads Spanning SNP[2] with SNP[3]}=\bullet)}$$

Fig 5 Approximating the $P(\bullet_2 \mid \bullet_3)$ conditional with inspiration from a Naive Bayes text classifier. Such pairwise conditionals are far cheaper to calculate.

Note that we are not interested in the identity or whole genomes of species that secrete these enzymes but rather the set of all haplotypes for an enzyme of interest.

Results

Initial testing of our approach was performed by randomly generating some number of fixed length DNA sequences to be used as mock haplotypes. Sets of short reads were derived from these random sequences to create trivial synthetic metahaplomes.

We also recovered haplotypes from a pair of synthetic metahaplomes consisting of short reads generated from five real variants of ***DHFR*** and ***AIMP1*** genes:

#Haplotypes	#SNPs	Max	Average	Min
3	10	100.00	96.33	60.00
	50	100.00	89.87	58.00
	250	100.00	76.53	50.80
10	10	100.00	87.20	50.00
	50	100.00	66.40	40.00
	250	99.60	49.94	32.40
25	10	100.00	73.92	40.00
	50	88.00	50.55	34.00
	250	58.80	37.90	29.20

Fig 6 ▲ Max, average and min recovery rates of our approach on a set of trivially generated metahaplomes with defined haplotypes and variants.
▼ Average recovery rates of our approach from synthetic metahaplomes each containing five real variants of either ***DHFR*** or ***AIMP1*** genes.

	Average Recovery Rate by Haplotype (%)				
Gene	H1	H2	H3	H4	H5
DHFR	92	94	79	73	76
AIMP1	97	96	97	92	60

Additionally, we obtained a 70% average recovery rate across 71 unique haplotypes in a metahaplome for a segment of the Influenza A viral genome. A haplotype in this data set, consisting of 264 variants, was recovered with 99% accuracy.

Why does this matter?

For the first time, we have shown it is possible to computationally extract variants of real genes from a metagenomes consisting of short sequenced reads.

What's next?

Use recovered haplotypes as primers for PCR of metagenomic DNA. Validate with identity between recoveries and single molecule sequences of amplicons.

Advances in the recovery of haplotypes from the metagenome dx.doi.org/10.1101/067215

msn@aber.ac.uk @samstudio8 samnicholls.net